By DAVID C. WEBER

A Quagmire of Scientific Literature?

EVER SINCE JULY 1945, when Vannevar Bush described the quandary of scientists who are swamped by the literature of their field, men working in pure science or technology have been worrying about bibliographical control over the flood of their publications which threatens to interrupt their own research.[1] John E. Burchard, writing four years after Bush, thought that the sheer bulk of published writings and the difficulties of quick and explicit accessibility were causing a literary "Waterloo of Science."[2] In 1953, Maierson and Howell stated that "for a number of years it has been apparent that conventional methods of indexing and classifying technical literature can no longer cope with the ever increasing flood. It is frequently more economical to repeat work of the past than to search the technical literature for the desired item. . . . "[3] And Mitchell clearly argued that "the tremendous increase in the volume of technical literature of all kinds and fields is presenting the librarian with an almost impossible reference task. The sheer volume of these documents is creating a filing problem of the first magnitude. When this volume is combined with the fact that many documents cut across classification lines, the problem of providing reference bibliographies is made that much more difficult."[4] Librarians as a group have been slow to realize that scientists are truly worried about their literature situation.

An analysis of this literature problem shows that in the last fifteen years the scientist has become a publisher in similar quantity to the humanists and social scientists of the last several centuries; and, in the field of science, the unit needing classification and housing and retrieval "has changed from macroscopic masses embodied in books to microscopic units embodied in articles."[5] A comparison of publishing method in different disciplines may reveal the cause of the scientists' dilemma. In the humanities and social sciences, publication is primarily divided between periodicals, which describe the results of new research, and monographs, which provide the more fully documented statements. For both of these, there is adequate listing and suitable indexing. In science, on the other hand, the publishing scheme is a complex one made up of the technical report, the pre-print, the periodical, and finally the monograph. There is little control bibliographically over the technical report, none over the pre-print, and only delayed control over the periodical. However, when the scientist is asked his information-gathering habits, he replies as follows, in this approximate order: his direct sources are advanced publications, research periodicals, technical reports, and handbooks, and his indirect sources are conversations, regular perusal of periodicals, references cited in books and papers, abstracts and indexes.[6]

[1] Vannevar Bush, "As We May Think," *Atlantic Monthly*, CLXXVI (1945), 101-08.
[2] John E. Burchard, "The Waterloo of Science," *Revue de la Documentation*, XVI (1949), 94-97.
[3] Alvin T. Maierson and W. W. Howell, "Application of Standard Business Machine Punched-card Equipment to Metallurgical Literature References," *American Documentation*, IV (1953), 3.
[4] Herbert F. Mitchell, Jr., "The Use of the UNIVAC Fac-tronic System in the Library Reference Field," *American Documentation*, IV (1953), 16.

Mr. Weber is assistant to the director, Harvard University Library.

[5] S. R. Ranganathan, *Philosophy of Library Classification* (Copenhagen: Munksgaard, 1951), p. 13.
[6] Saul Herner, in a paper entitled "The Information Gathering Habits of Johns Hopkins Scientists" which was reported by Marjorie R. Hyslop in her "Documentalists Consider Machine Techniques," *Special Libraries*, XLIV (1953), 197-98.

It should be evident, therefore, that unified bibliographical control over this variety of publishing forms is really the problem, and the difficulty is not caused by any form of informational freakishness which should force librarians or scientists to turn to machine storage in order to gain access to the material they need. It is all too often that the scientist or documentation expert starts his argument with the thesis that scientific literature is flooding the laboratories and proceeds to argue for the development of the Memex, Ultrafax, Rapid Selector, Avakian's AMFIS, Minicard, and other complex and expensive devices for storage and retrieval of information.

It can often be seen, through hindsight, that a problem has not been tackled by a slight adjustment but by a wholly new process or device which many times proves less suitable than the old process in its greatest development. Battelle Memorial Institute reported one typical instance in a recent evaluation of techniques commonly used for literature collection and analysis. "It became quite evident during preliminary investigations that the old-fashioned manual systems had not previously been thoroughly evaluated and that these techniques, thought to be outdated, seemed not to have been fully exploited in the past. It was concluded that the time had come for re-evaluating manual systems or combinations of manual-machine methods before proceeding exclusively to the evaluation and development of machine systems."[7] The result at Battelle was a completely manual system.

The handling of difficult collections of materials, be they pamphlets, reprints, serials, documents, or monographs, has been the long-standing business of the library profession. If the librarian in the disciplines of pure science and technology professes inability to handle these materials and produce the information desired by the scientists, it may well be that the librarian's approach is wrong, that the library is understaffed, or that there is not enough money put into the bibliographical apparatus—an expense which is not so glamorous a way to spend money as would be some unorthodox machine. To put it another way, if the indexing and abstracting services in science do not provide the information which is needed, the librarian should make every effort to do this listing and indexing for reports, pre-prints, or periodical articles, whenever needed by his clients, just as he now does for monographic materials. It is a simple problem, and the solutions are also simple, though they may be moderately expensive.

Subject analysis of material—and its corollary, the location of material from a subject approach—is a separate and distinct problem from that of author and title listing. The latter is but a temporal problem needing concerted attention. But the subject approach to one's library is ordinarily fragmentary indeed, as compared to the relative comprehensiveness for the author approach, so it should *ipso facto* be part of a system of indexes designed to reveal what exists anywhere in print on the particular subject of concern. However, take the scientist who professes interest in the subject content of only his own library, perhaps because he can assume his library is all but comprehensive within his interests. Even here, library methods of an orthodox type can do practically everything a machine can, and can generally do it faster. Shaw has said that, "depending upon the type of search, it is even doubtful whether the fastest electronic machine that we can postulate will ever be able to search for a series of *author* entries as rapidly or as economically as . . . can be done in a conventional card catalog." And he goes on to say that "when large files have to be maintained and when they have to be

[7] Robert W. Gibson, Jr., and Ben-Ami Lipetz, "New Look in Manual Methods," *Special Libraries*, XLVII (1956), 108.

searched repeatedly for *subject* information, great reduction in space requirements and in searching time and in copying time may be achieved by mechanization."[8]

Even this qualified statement, by a person who is adept at machine application, suggests more for the machine than should be expected. The most important factor which is usually overlooked is that the machines contribute substantially only to the consumption end, not the production end; because human cataloging or encoding is the essential preliminary to any mechanized storage and consultation. Vannevar Bush is at his most imaginative when he outlines how machines might hurdle this biggest of problems: "When the user is building a trail, he names it, inserts the name in his code book, and taps it out on his keyboard."[9] Note that the human being must "name" the subject before the machine can store and return it for use; machines cannot yet replace traditional library methods in this analysis. And even on the consumption end, Dr. Bush reminds us that "the prime action of use is selection, and here [machines] are halting indeed."[10]

Let us turn to more minute concerns. Discussion as to the relative merits of card catalogs and storage machine frequently boils down to two capacities: high subject specificity and multiple subject approach. Specificity refers to subject access at the particular level rather than the general. It is one thing to put a book on female cat diseases under a subject heading MEDICINE. It is more specific to put it under the heading MEDICINE—ANIMALS, or, even more specific, under MEDICINE—CATS—FEMALE. Although librarians have always aimed at placing a book under its most specific heading, it has been understood that this would never be taken to extremes. On the other hand, scientists want headings that regularly place the information under the most specific heading possible. Taxonomic classification, based on family relationships, would theoretically satisfy everyone; but neither for machines nor for a classed catalog has a universally acceptable taxonomic classification for the entire range of knowledge been developed. Under any condition, therefore, the card catalog can do as well as the machine on specificity.

As for multiple subject approach, classification of books on the shelf provides single access, and this does not suffice for adequate subject approach in the sciences, nor even in the humanities and social sicences. However, card catalogs, and particularly classed card catalogs, can satisfy this need. A book that is listed in the catalog under the headings CATS and VETERINARY MEDICINE and ANIMAL DISEASE will be given three approaches. Here again, the card catalog is theoretically as versatile as the machine.

To see where machines run into their basic trouble, one has only to consider the mathematical structure of language. Language, as analyzed by symbolic logic, presents extreme complications to the coding process and the subsequent retrieval; for every language has built-in entropy (electronic's "noise"), in phonetics, semantics, inflection, and syntactical construction. However, definition in terms of probabilities goes far to point out a solution, even allowing full weight to redundancy (whether it is the "K" of *key* and the "K" of *cool,* or "page" as a *messenger* or *leaf* of a book) ; but it is still only a theory, which will not come to practical application for many years. In his discussion of machine translation which involves coding followed by decoding, Whatmough explains this small but as yet unsurmounted barrier:

A human translator has the necessary circulatory pathways established already as patterns of neural activity by virtue of being

[8] Ralph Shaw, "Mechanical and Electronic Aids for Bibliography," *Library Trends,* II (1954), 530-31. Italics mine.
[9] Bush, *op. cit.,* 107.
[10] *Ibid.,* 105.

bilingual. It appears likely, simply in terms of regional examination of the human living brain and its functioning that speech and "thought" are very much connectible. Language to a tremendous extent is a matter of habit—if it were not, communication would be impossible; but the areas of association on the basis of which most of our linguistic and non-linguistic behavior is to be accounted for, the socio-personal areas, are so closely linked, that cerebration, if done symbolically, with both the outside universe and inner "experience" as a unified frame of reference, is done with linguistic symbolism, or at least within a system of operations based on linguistic symbolism.[11]

Machines imitate the human brain which is based on the neuron's binary action and which handles morphemes (words or independently significant parts) rather than phonemes (parts of words which are minimum speech sounds).

But, and this is the crux of the matter, the machine must now be provided with a statistical distribution law for the relative frequency of occurrence of the units and constructions of language, the "circulatory pathways" using "linguistic symbolism," in order for it to be an information system independent of restrictions of subject matter, size of vocabulary, human pre-editing or post-editing, and the amount of text. Such a law is not yet within sight. Taube and his associates found that a "dictionary of associations" would be necessary to solve many of the semantic problems still faced by their system of coordinate indexing.[12] And, most recently, Perry and his associates have spent years working on machine literature searching before finding that the coding system for machines would have to use symbolism for "semantic factors" and "analytic relationships" and that a "code dictionary" would have to be constructed so as to deal with language problems.[13]

The conclusion to be drawn is that the use of machines for storage and retrieval of information is likely to be practicable only through a man-machine partnership, and is not going to be commonly feasible for many years to come. If financial costs can be left out of the question, and if specificity and multiple approach are not critical determinants, under what conditions may storage machines be superior to the card catalog? It is here contended that the machine will be the better choice only when all of the following conditions prevail:

1. A single subject is being covered.
2. There is a high concentration of publications in this subject area.
3. There is a continuing high intake rate of publications.
4. Adequate subject access is unavailable in published form.
5. Use is made by people having several different approaches or uses in mind.
6. There is high urgency in the location of every pertinent publication.

In such a case, there is a probability that some unorthodox method of storing and retrieving information may be required. (The Uniterm system of coordinate indexing seems suitable only when the above conditions apply and when the collection indexed is not to reach 100,000 items.) Shera says that the use of machines "seems likely to be limited to the more complex problems of bibilographical searching, and therefore, they may not be applicable to the entire range of bibliothecal operations."[14]

It is nevertheless unquestioned that libraries in science and technology must improve in order to cope with the growth of their diverse literature. Comprehen-

(Continued on page 118)

[11] Joshua Whatmough, *Language, a Modern Synthesis* (New York: St. Martin's Press, 1956), 213-14.

[12] See page 7 and *passim* in Mortimer Taube and Associates, "Storage and Retrieval of Information by Means of the Association of Ideas," *American Documentation*, VI (1955).

[13] James W. Perry, Allen Kent and Madeline M. Berry, *Machine Literature Searching* (New York: Interscience Publishers, 1956), p. 84.

[14] Jesse H. Shera, "Effect of Machine Methods on the Organization of Knowledge," *American Documentation*, III (1952), 16.

VI. Inquiries

Search circulation file for book card.	Same as A	Search circulation file for charge card.
If no card is found, book is in the library.	Same as A	Same as A.
If card is found, book is out.	Same as A	When card is found, check transaction number on check list. If the transaction number is checked off on the list, book has been returned, and the charge card should be discarded. If the number is not checked off, book is still out.

VII. Book Collections Checked Out and Returned in One Parcel
(Reserve room, departments, binding, class room, etc.)

Books must be discharged individually.	Same as A	By use of specially coded cards, the sorting key can be used to discharge the whole collection at once. This method can be used also for taking inventory of books loaned to a special collection.

A Quagmire of Scientific Literature?

(Continued from page 106)

sive current subject bibliographies are a primary need. Tauber has stated that "it is almost certain that more selective subject catalogs and more extensively used subject bibliographies will characterize subject analysis in the immediate future."[15] A secondary need is for comprehensive indexing of serial publications, where the situation is distinctly unsatisfactory. Librarians have been ineffectual in eliminating wasteful overlapping of services and in obtaining inclusive indexing; this is a critical situation into which must be put much more effort.[16] It is logical to expect that a great increase in extremely brief subject entries, arranged in chronological order, will characterize the future subject indexes to scientific materials—with the older material being indexed merely by an author file, and with subject cards thrown out after a period of time.

It can be said with complete assurance that scientific libraries have somewhat different problems from libraries in other disciplines, that they are still far from having satisfactory bibliographical control over scientific literature, and that existing library methods if fully exploited can bring firm ground out of the quagmire that now seems to be threatening.

[15] Maurice F. Tauber and Associates, *Technical Services in Libraries* (New York: Columbia University Press, 1954), p. 175.

[16] On the indexing situation, see Verner W. Clapp and Kathrine O. Murra, "The Improvement of Bibliographic Organization," *Library Quarterly*, XXV (1955), 107.